DOCUMENT RESUME

ED 108 631 IR 002 148

AUTHOR Porch, Ann; Lang, Pat
TITLE Design Document: Content Module; L.A.P. Version I.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
REPORT NO SWRL-TN-5-72-35
PUB DATE 26 May 72
NOTE 12p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Computer Programs; Content Analysis; Data Processing; Design; Electronic Data Processing; Item Analysis; Language Patterns; *Language Research; Languages; Language Styles; Language Usage; *Morphology (Languages); *Specifications; *Systems Development
IDENTIFIERS Analytical Computer Packages; Computer Software Documentation; Content Module; *Language Analysis Package; LAP

ABSTRACT

A series of computer programs and routines designed to assist researchers in the analysis of language usage was developed by the Southwest Regional Laboratory (SWRL). This document is one of a series that describes design specifications for the individual modules which comprise the Language Analysis Package (LAP). The Content Module functions as a semantic content analysis module by allowing the user to construct any number of dictionary files using phrases and/or single words. Each dictionary file will represent a set of user-defined semantic categories. The program will "score" the input text by matching it against the dictionary and it will give the user the total number of different categories a particular word or phrase falls into, as well as the categories into which it falls. Program characteristics, data file specifications, and computer file layout are provided. (Author/DGC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED108631

IR

SWRL

SOUTHWEST REGIONAL LABORATORY
TECHNICAL NOTE

DATE: May 26, 1972

NO: TN 5-72-35

TITLE: DESIGN DOCUMENT: CONTENT MODULE - L.A.P. VERSION I

AUTHOR: Ann Porch and Pat Lang

ABSTRACT

This is one of a series of technical design specifications for individual modules of the Language Analysis Package (L.A.P.).

The Content Module will allow the user to specify semantic categories, and to obtain an analysis of the input text in terms of those categories.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

IR 002 148

This document is intended for internal staff distribution and use. Permission to reprint or quote from this working document, wholly or in part, should be obtained from SWRL, 11300 La Cienega Boulevard, Inglewood, California, 90304

DESIGN DOCUMENT: CONTENT MODULE - L.A.P. VERSION I

This document is one of a series of programming design specifications for individual modules of the Language Analysis Package (L.A.P.)[1]. The section of the system design to which it is related is 6.7.0.

Program Objective

Version I of the Content Module[2] will function as a semantic content analysis module by allowing the user to construct any number of sorted dictionary files using phrases and/or single words. Each dictionary file will represent a set of user-defined semantic categories. The program will "score" the input text by matching it against the dictionary and give the user the total number of different categories a particular word or phrase falls into, as well as the categories in which it falls

---

1. See Porch, Ann. TM 5-72-06 "Language Analysis Package (L.A.P.) System Design (Version I)" for an overview of the package.
2. A modification of an existing program will be used for the Version I Content Module. The program is SCORTXT, developed by Gerald Fisher of the University of Connecticut. Full documentation on the program may be found in Fisher, Gerald. "The SCORTXT Program for the Analysis of Natural Language", University of Connecticut, Bureau of Educational Research.

Constraints and Limitations

- When the text is scanned for dictionary matches, only the longest string will be matched. For example, if the dictionary contains both "not very much" and "not very," the text phrase "not very much" will be counted as matching only "not very much".

- No punctuation (except apostrophes and hyphens) may be included in dictionary phrases. Texts longer than 1,500 words must be broken into sub-texts if high-frequency words are category entries.

Options and Defaults

Options and defaults for the Content Module will be as follows:

- Print text in original form (Default = no print)
- Print text in array form (Default = no print)
- Print sorted dictionaries (Default = no print)
- Print a reduced text (Default = no print)
- Print an item analysis of each category (Default = print)
- User specified word length for text in array form (Default = 16)
- User specified input record margins for text (Default = 1,72)

Data File Specifications

Input files for Version I must follow the ordering shown in Appendix A. Dictionary input will consist of one or more dictionaries.

Output for Version I will consist of punched or printed output as shown in Appendix B.

Significant Algorithms

Category Indicator Algorithm:

After a dictionary is read in and sorted internally into alphabetical sequence, a bit string will be created for each dictionary entry with on-off indicators for each category represented. Thus if the dictionary has 1,000 entries which fall variously under five categories, then for each of the 1,000 entries a bit string of length five will be created with 1's in each category position to which the entry belongs. The sorted dictionary and the dictionary bit strings are added to the file DICT.

Significant Variables

There are no variables of special significance associated with the Content Module.

Error and Other Messages

The following messages are printed out by the Content Module:

- "Dictionary Not Found" if there is no dictionary associated with the input for a particular run.

- "End of Job" if the run terminates normally when there are no further texts to be read.

Called by and/or Call

The Content Module is called only by the Control Module.

The Content Module will call the following internal subroutines:

- ARRAY
- DINIT

- INDXDLM
- ITEM
- LSTAT
- MAKEDIC
- SORT
- PHRASE
- INTABLE

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

TITLE: Content Module DATE: May 8, 1972

FILE ID: Input

PROGRAMMER:

RECORDING MODE:
FIXED ☐
LENGTH ___
VARIABLE ☐
MAX ___
MIN ___

BLOCKING FACTOR:
BLOCK CONTAINS
___ RECORDS

REMARKS:
Appendix A Input 1 Text Input

CARD LAYOUT ☐

TAPE
RECORD LAYOUT ☐

FORMAT LEGEND:

△ = Blank	P = Packed Dec.
A = Alpha	O = Octal
N = Numeric	H = Hexadecimal
λ = Alpha/Num.	B = Binary

RECORD I.D. Text Header (one per text)	Text ID	User Comments	
FORMAT			
POSITIONS	5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100		
☐ CONTINUED RECORD I.D. Text (as many as needed)	text punched (029) within user specified card margins		
FORMAT			
POSITIONS	5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100		
☐ CONTINUED RECORD I.D. End of Text (more texts may follow)	**		
FORMAT			
POSITIONS	5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100		
☐ CONTINUED RECORD I.D. End of All Texts (Dictionary Selection Input must follow)	$$		
FORMAT			

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

RECORDING MODE:
FIXED ☐
LENGTH ____
VARIABLE ☐
MAX ____
MIN ____
BLOCKING FACTOR:
BLOCK CONTAINS
____ RECORDS

REMARKS:
Dictionary Input Input 2

TITLE: Content Module DATE: ____
FILE ID: Input
PROGRAMMER: ____

CARD LAYOUT ☐
TAPE
RECORD LAYOUT ☐

FORMAT LEGEND:

△ = Blank	P = Packed Dec.
A = Alpha	O = Octal
N = Numeric	H = Hexadecimal
X = Alpha/Num.	B = Binary

RECORD I.D.
Dictionary Header
(one per dictionary)

Dictionary Name

← Maximum number of characters in any entry

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
Dictionary Entry
(as many as are needed)

Dictionary Entry

category no.

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
End of Dictionary
(more dictionaries may follow)

**

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
End of All Dictionaries
(Text Input must follow)

$$

FORMAT

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

TITLE: Content Module DATE: May 8, 1972

FILE ID: Input

PROGRAMMER:

RECORDING MODE:
FIXED ☐
LENGTH ___
VARIABLE ☐
MAX ___
MIN ___

BLOCKING FACTOR:
BLOCK CONTAINS
___ RECORDS

REMARKS:

Dictionary Selection Input Input 3

CARD LAYOUT ☐

TAPE
RECORD LAYOUT ☐

FORMAT LEGEND:

△ = Blank	P = Packed Dec.
A = Alpha	O = Octal
N = Numeric	H = Hexadecimal
X = Alpha/Num.	B = Binary

RECORD I.D.
Dictionary Selection
(as many as needed)

Dictionary Name

Dictionary ID num

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
End of All Dictionary
Selection (last input
record)

$$

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

6 RECORD I.D.

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.

FORMAT

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

TITLE: Content Module DATE: May 8, 1972

FILE ID: OUTPUT

PROGRAMMER:

RECORDING MODE:
FIXED ☐
LENGTH ___
VARIABLE ☐
MAX ___
MIN ___

BLOCKING FACTOR:
BLOCK CONTAINS
___ RECORDS

REMARKS:

Appendix B Output 1
Score Result Output

CARD LAYOUT ☐

TAPE
RECORD LAYOUT ☐

FORMAT LEGEND:

△ = Blank	P = Packed Dec.
A = Alpha	O = Octal
N = Numeric	H = Hexadecimal
X = Alpha/Num.	B = Binary

RECORD I.D.
Score Result (as many as needed - one output record per 18 categories)

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

5	10	15	20	55	60	80
Text ID	category 1 score	category 2 score	category 3 score	category 4 score	category 5 score	number of categories is user controlled (up to 18 per record)	category 18 score	Dictionary ID no.

☐ CONTINUED

RECORD I.D.
Text in Array Form (optional)

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

5	10	15	80
Text ID	Total Number of words	Word Length of longest word	Text words printed in array form (number of columns each = 80/word length of longest word.)

10

☐ CONTINUED

RECORD I.D.
Dictionary (one per dictionary)

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

60	65	70	75	80
Dictionary Name	No. of entries	Maximum of Entry length	No. of categories	Entry specification 0=single word 1=phrases

☐ CONTINUED

RECORD I.D.
Dictionary (Cont.) (as many as there are entries in dictionary)

FORMAT

Dictionary Entry + bit string representing categories to which it belongs (Lengths of entry and bit string vary.)

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

RECORDING MODE:
FIXED ☐
LENGTH ____
VARIABLE ☐
MAX ____
MIN ____
BLOCKING FACTOR:
BLOCK CONTAINS ____ RECORDS

REMARKS:
Output 2

TITLE: Content Module DATE: May 8, 1972

FILE ID: OUTPUT

PROGRAMMER:

CARD LAYOUT ☐

TAPE RECORD LAYOUT ☐

FORMAT LEGEND:
△ = Blank P = Packed Dec.
A = Alpha O = Octal
N = Numeric H = Hexadecimal
X = Alpha/Num. B = Binary

RECORD I.D. Reduced Text (optional)	Dictionary Name	Text ID	Number of words in text	Number of categories	
FORMAT					
POSITIONS	5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100				

☐ CONTINUED

RECORD I.D.
Reduced Text (con't)

Reduced Text (bit string for each word in text, indicating all those categories to which a word belongs. The string will be filled in with zeros for a word not found in any category. If there is a match on a multi-word phrase, then the first word's bit string will indicate all the categories to which the phrase belongs, and the bit strings for all other words in the phrase will be filled with ones)

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

11 ☐ CONTINUED

RECORD I.D.
Item Analysis (optional)

"ITEM ANALYSIS FOR DICTIONARY 'NAME'"

FORMAT

POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED

RECORD I.D.
Item Analysis (con't)

"COUNTS FOR CATEGORY 'NUMBER'"

FORMAT

SWRI

COMPUTER SYSTEMS
FILE/RECORD LAYOUT

TITLE: Content Module DATE: May 8, 1972

FILE ID: OUTPUT

PROGRAMMER:

RECORDING MODE:
FIXED ☐
LENGTH ____
VARIABLE ☐
MAX ____
MIN ____

BLOCKING FACTOR:
BLOCK CONTAINS ____ RECORDS

REMARKS

Output 3

CARD LAYOUT ☐

TAPE RECORD LAYOUT ☐

FORMAT LEGEND:

△ = Blank	P = Packed Dec.
A = Alpha	O = Octal
N = Numeric	H = Hexadecimal
X = Alpha/Num.	B = Binary

RECORD I.D.
Item Analysis (con't)

"CATEGORY NAME" "CATEGORY COUNT"
or "NO COUNT FOR THIS CATEGORY"

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED
RECORD I.D.

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED
RECORD I.D.

FORMAT
POSITIONS 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100

☐ CONTINUED
RECORD I.D.

FORMAT